Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 78 (2005) 19–29

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Explaining a presence of groups in analytical data in terms of original variables

M. Daszykowski[a,b], I. Stanimirova[a], B. Walczak[a,b,*], D. Coomans[a,c]

[a]*ChemoAC, Vrije Universiteit Brussel, FABI, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b]*Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland*
[c]*Statistics and Intelligent Data Analysis Group, School of Mathematical and Physical Sciences, James Cook University, Townsville Q4814, Australia*

Received 19 October 2004; received in revised form 1 December 2004; accepted 6 December 2004
Available online 8 February 2005

Abstract

This manuscript shows the usefulness of Projection Pursuit (PP) and Multivariate Regression Trees (MRT) for analytical data exploration. Additionally, features of Projection Pursuit and kurtosis as a projection index are presented. The ability of Projection Pursuit to discover groups in the data is compared to classical Principal Component Analysis (PCA). Moreover, it is also demonstrated how the presence of groups in the data can be explained in terms of explanatory variables with the aid of Projection Pursuit and Multivariate Regression Trees. Neither Projection Pursuit nor Multivariate Regression Trees are commonly used for exploring chemical data however, they are able to enrich to a high extent the interpretation.

Keywords: Exploratory analysis; CART; MRT; Projection Pursuit; Projection index; Kurtosis

1. Introduction

In analytical chemistry, the modern instruments can assemble easily a large portion of information about studied samples. The exploration aspect of the collected data plays a very fundamental role in an overall experimental process and it is one of its latest stages. Usually, the chemical data are with a complex structure and multidimensional. Typical chemical data contain many samples described by various physico-chemical properties. Particularly in some cases, hundreds or even thousands of variables constitute chromatographic or spectral data. Thus, to point out any conclusions about the similarities among samples, a data visualization is required. The visualization and/or compression of the data can be done using chemometrical approaches. Principal Component Analysis (PCA) is a widely used chemometrical approach for this purpose, and is also the best-known projection technique [1,2]. The goal of PCA is to project multidimensional data onto a space of a few orthogonal variables called Principal Components (PCs). PCs are a linear combination of the original data variables, and they are obtained by maximizing the data variance. Each of the new PCs describes a part of the data variance not modeled by its predecessors. PCs encountering for the most of the data variance can be used for a data structure visualization. Frequently, the projection of objects on the space spanned by the first two or three PCs uncovers a specific data structure (clusters and/or outlying objects). The clustering techniques, which group the data, do not provide any information about original variables that cause the grouping. Discovering groups in the data can additionally be done with the aid of Projection Pursuit (PP), not commonly applied in the field of chemistry. Projection Pursuit (PP) aims to find 'interesting', low-dimensional projections (one-, two- or three-dimensional projections), which reveal a unique data structure [3] (clusters and/or outliers). These projections are found by optimizing a

* Corresponding author. Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland. Tel.: +48 32 359 12 46; fax: +48 32 259 99 78.
E-mail address: beata@us.edu.pl (B. Walczak).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2004.12.002

certain projection index, describing the heterogeneity of the data. The objective of the method (looking for clusters or outliers) depends on the selected projection index. Contrary to clustering methods, PP can discover groups in the data subspaces, which is certainly an advantage of PP, compared to the clustering methods. The latter ones cluster the objects in the whole data space. Unfortunately, Projection Pursuit Features (PPFs) are also a linear combination of original data variables as PCs. Of course, PP loadings can be used for interpretation of the PP features in terms of explanatory variables, but the obtained information about the variables' importance is not evident. The same problem may occur while interpreting PCA loadings, if a few loadings have similar values. Therefore in order to facilitate interpretation of groups revealed on the PPF projections in terms of explanatory variables, Multivariate Regression Trees (MRT) can be used. Multivariate Regression Trees are an extension of the regression trees, proposed by Breiman and Friedman [4–6], which handle more than one response variable. A two-step procedure proposed in this manuscript, PP followed by MRT, enables challenging two aspects of the data exploration simultaneously—discovering groups in the data, and finding the reasons of the groups' presence by analyzing the discrimination power of the individual variables. Interpretation of the groups in the data in terms of the individual variables can be also done with other techniques, for instance, Linear Discriminant Hierarchical Clustering, proposed by Marengo et al. [7].

2. Theory

2.1. Projection Pursuit

The aim of Projection Pursuit, PP, is to find interesting low-dimensional projections within the data space emphasizing the clustering tendency or outliers [3]. PP can be considered as a generalization of classical PCA, where the data variance is presented as a projection index. From a practical point of view, an interesting projection reveals a clustering tendency of the data and/or highlights the presence of outliers (objects with atypical properties). In order to describe the projection, i.e. whether it is interesting or not, different projection indices have been developed. The least interesting one-dimensional projection has a normal distribution [8]. For this reason, the majority of the projection indices are designed to be sensitive to any deviation from the normal distribution. The projection index is a continuous function of the data distribution and is uniquely minimized by the normal data distribution. One of the most popular projection indices is entropy, expressing the lack of organization or ordered structure in the data [8].

PP can also be viewed as an optimization task of maximizing a projection index within the data space. The maximization procedure within PP relies on the principles of hill-climbing which results in finding different local maxima of the projection index [9]. The history of PP can be traced to the early fifties, when Roy [10] has described a preliminary idea of PP. Later on, the concept of low-dimensional projections showing a unique structure of the data has been developed by Kruskal [11]. The successful application of PP is due to Friedman and Tukey [3], who also coined the term "projection pursuit".

There are different approaches searching the space of the variables to find interesting low-dimensional projections. In Sequential Projection Pursuit of Guo et al. [12] the projections are found by Genetic Algorithm. In our studies, the sequential algorithm of Croux et al. [13] is used. At the very beginning of PP, the data set is preprocessed such that the mean (the 1st central moment) and the variance (the 2nd central moment) of each variable is equal to 0 and 1, respectively. In the PP terminology this step is known as 'sphering' or also 'whitening'. Such a data transformation removes the differences in the variables' range and stretches the data in each direction of the space equally [14].

2.2. Kurtosis as a projection index

The most attention is drawn to projection indices, which can be derived from higher moments of the data distribution, mostly due to computational aspects [15,16]. For instance, the entropy can be approximated by higher-order cumulants [16,17]. The higher moments of the data distribution are skewness and kurtosis. In general, skewness describes asymmetry of the data distribution, whereas kurtosis measures departure from the normal distribution. Pena demonstrated in Ref. [18] that kurtosis has its maximal value when the projection of the data on a certain direction contains several groups of objects of different size or it contains outliers. Small values of kurtosis are typical for bimodal projections, i.e., with two well-separated groups of similar size [19]. For these reasons, kurtosis seems to be a well-suited projection index and capable to emphasize projections with a high clustering tendency and/or with outliers. Kurtosis is affine invariant, which means that it remains unchanged under certain class of data transformation, and additionally it satisfies Huber's conditions for a good projection index described in Ref. [8]. Another advantage of this projection index is that it can be computed fast. Kurtosis of the centered projection, $\boldsymbol{x}$, is defined as the fourth central moment, $\boldsymbol{x}(4)$, divided by the fourth power of standard deviation (the second central moment), $\boldsymbol{x}(2)$:

$$kurt(\boldsymbol{x}) = \frac{\boldsymbol{x}(4)}{\boldsymbol{x}(2)^4} \tag{1}$$

It should be born in mind that the number of considered data samples has a crucial impact on determining the above projection index as well as any statistics. Generally, the more objects in the data, the more reliable estimates of mean, standard deviation and kurtosis, can be obtained.